Yong Yang[1]
Theodore W. Thannhauser[1]
Li Li[1]
Sheng Zhang[2]

[1]US Plant, Soil and Nutrition Laboratory, USDA-ARS, Ithaca, NY, USA
[2]Proteomics and Mass Spectrometry Core Facility, Cornell University, Ithaca, NY, USA

Received May 25, 2006
Revised November 30, 2006
Accepted January 8, 2007

Research Article

Development of an integrated approach for evaluation of 2-D gel image analysis: Impact of multiple proteins in single spots on comparative proteomics in conventional 2-D gel/MALDI workflow

With 2-D gel mapping, it is often observed that essentially identical proteins migrate to different positions in the gel, while some seemingly well-resolved protein spots consist of multiple proteins. These observations can undermine the validity of gel-based comparative proteomic studies. Through a comparison of protein identifications using direct MALDI-TOF/TOF and LC-ESI-MS/MS analyses of 2-D gel separated proteins from cauliflower florets, we have developed an integrated approach to improve the accuracy and reliability of comparative 2-D electrophoresis. From 46 spots of interest, we identified 51 proteins by MALDI-TOF/TOF analysis and 108 proteins by LC-ESI-MS/MS. The results indicate that 75% of the analyzed spots contained multiple proteins. A comparison of hit rank for protein identifications showed that 37 out of 43 spots identified by MALDI matched the top-ranked hit from the ESI-MS/MS. By using the exponentially modified protein abundance index (emPAI) to determine the abundance of the individual component proteins for the spots containing multiple proteins, we found that the top-hit proteins from 40 out of 43 spots identified by MALDI matched the most abundant proteins determined by LC-MS/MS. Furthermore, our 2-D-GeLC-MS/MS results show that the top-hit proteins in 44 identified spots contributed on average 81% of the spots' staining intensity. This is the first quantitative measurement of the average rate of false assignment for direct MALDI analysis of 2-D gel spots using a new integrated workflow (2-D gel imaging, "2-D GeLC-MS/MS", and emPAI analysis). Here, the new approach is proposed as an alternative to traditional gel-based quantitative proteomics studies.

Keywords:
Comparative proteomics analysis / 2-DE / LC-ESI-MS/MS / MALDI-TOF/TOF / Protein abundance index
DOI 10.1002/elps.200600524

1 Introduction

MS-based proteomics typically involves the large-scale identification, quantification, and characterization of proteins at the subcellular, cellular, tissue, or organism levels. As comprehensive genomic sequence information becomes available for an ever-increasing number of species, MS-based proteomics will continue to be an important tool for the analysis of biological systems and the exploration of complex protein functions and interactions. In addition, proteomic approaches can provide valuable insight into the functional implications of post-translational modifications (PTMs) [1, 2].

Current technologies used for proteomic studies are based on a variety of separation techniques followed by identification of the separated proteins and proteolytic peptides using MS [3, 4]. One popular front–end separation technique is high-resolution 2-D gel electrophoresis which is capable of resolving >2500 distinct protein spots in a single analysis [5, 6]. The protein spots are excised from the gel, digested, and identified by MS and subsequent database searching [6–9]. An alternative approach to 2-DE is liquid

Correspondence: Dr. Sheng Zhang, Proteomics and Mass Spectrometry Core Facility, 135 Biotechnology Building, Cornell University, Ithaca, NY 14853, USA
E-mail: sz14@cornell.edu
Fax: +1-607-254-4847

Abbreviations: **α-CHCA,** α-cyano-4-hydroxycinnamic acid; **emPAI,** exponentially modified protein abundance index; **FA,** formic acid; **IDA,** information-dependent acquisition; **PTM,** post-translational modification; **WT,** wild type

chromatographic separation of peptides coupled with ESI MS or tandem MS detection [10–12]. This approach can involve a single RP separation; however, more often it incorporates a 2-D separation where the first dimension is an ion exchange chromatography and the second is RP.

Among the traditional mass spectrometric methods available for identification of proteins separated by 2-D gel, PMF by MALDI-TOF MS [13] and MALDI-TOF/TOF MS [14], and ESI-MS/MS [15] have been used almost exclusively for this application. Due to its inherent simplicity and relatively high sample throughput, MALDI-TOF is commonly used as a screening method for direct analysis of gel-extracted peptides. Confidence in correct identification of the proteins by this approach can be increased through the inclusion of a tandem TOF (TOF/TOF) mass analysis [14]. For spots that fail to be identified in the initial screening process, an additional analysis using either LC-MALDI-MS/MS, or LC-ESI-MS/MS can be applied. By implementing an LC separation, it is possible to minimize ion suppression and achieve peptide enrichment to improve detection sensitivity and dynamic range. In particular, on-line capillary (Cap) or nano-LC interfaced with a tandem mass spectrometer operated in data-dependent MS/MS acquisition mode will make the second stage analysis more comprehensive and allows higher throughput [16].

Despite the increasing popularity of chromatographic alternatives, 2-DE is still the most widely used means of resolving complex protein samples [17]. 2-DE provides high-resolution protein separation and enables the identification of PTMs and proteolytic processing in a convenient "reference map" format. Furthermore, highly reproducible differential protein expression profiles can be obtained for protein extracts under different cellular states, particularly with the introduction of sophisticated image analysis software and the use of "difference gel electrophoresis" (DIGE) techniques [18, 19].

While there are many advantages of a 2-D gel approach, it does have some well-known limitations which have been detailed in several recent reviews [7, 20, 21]. The most significant limitation is that it is impossible to resolve an entire proteome in a single electrophoretic analysis. Large-format, high-performance gels are capable of resolving >2500 individual protein spots, but many genomes of interest (such as cauliflower) contain 10^4–10^5genes. These can be expected to produce 10^5–10^6 distinct proteins when splice variants, translation variants, and PTMs are taken into account. Thus, it is not surprising that many seemingly well-resolved protein spots isolated from 2-D gels are found to contain more than a single protein. This observation poses a significant challenge for the proper interpretation of comparative gel experiments in which changes in protein abundance are inferred from changes in "gel spot volumes" (staining intensities). Clearly it is not sufficient to know that a particular spot volume has increased/decreased by some amount judged to be statistically significant; it is also necessary to know how that change is distributed among the various proteins present in the spot.

One way to lessen this problem is by prefractionating the proteome such that the 2-D gels of individual fractions contain a sufficiently small number of proteins and hence comigration during 2-D electrophoresis would be unlikely. While such an approach can be expected to greatly improve the overall resolution of the experiment and decrease the number of gel spots found to contain multiple proteins, it would make it more difficult to compare protein abundance data between the different fractions.

An alternative approach is to use a broad p*I* range 2-D gel as the front–end separation strategy and follow this with an LC-MS/MS analysis of the digested gel spots. Such a procedure would eliminate the need to normalize abundances between fractions, but would require a methodology to determine the relative abundance of proteins present in a digested gel spot from the mass spectral data alone. Recently, Ishihama *et al.* [22] demonstrated a method for determining the amount of each protein in a proteomics sample by using the number of detected peptides *per* protein to create an "exponentially modified protein abundance index" (emPAI).

As part of a study to understand the mechanisms involved in the accumulation of carotenoids in a cauliflower mutant [23], we applied 2-DE to separate proteins extracted from the edible curd of both homozygous wild-type (WT) cauliflower and the orange-colored homozygous mutant (*Or*), to identify proteins associated with the accumulation of carotenoids. Here, we report the comparison of protein identifications made by direct MALDI-TOF/TOF and CapLC-ESI-MS/MS analysis. This work compared several aspects of the MS analysis approaches including the number of identified proteins, their matched peptides, sequence coverage, protein-hit ranks, and the utility of combining the two methods. The initial comparison led us to develop an integrated workflow to quantitatively evaluate the abundance of each protein identified within a single gel spot. This novel approach integrates 2-D gel-based GeLC-MS/MS analysis [24] with an empirically derived tool (emPAI) [22] to correctly distribute the change in expression determined by gel staining and image analysis to each of the protein constituents in the spot. The impact of multiple proteins in single spots on the results of comparative image analysis and data interpretation is discussed.

2 Materials and methods

2.1 Materials

Deionized water from a Milli-Q ultrapure water system (Millipore, Bedford, MA, USA) was used and sequence-grade ACN and acetone were purchased from Fisher Scientific (Fair Lawn, NJ, USA). Ammonium phosphate, α-cyano-4-hydroxycinnamic acid (α-CHCA), ammonium bicarbonate, iodoacetamide (IAM), DTT, BSA, and lactoferrin were from Sigma (St. Louis, MO, USA). Modified porcine trypsin was purchased from Promega (Madison,

WI, USA). Urea, SDS, and all the other chemical reagents, unless otherwise noted, were obtained from Aldrich (Milwaukee, WI, USA).

2.2 Preparation of cauliflower protein extracts

Two grams of frozen curd material of both WT cauliflower and the orange-colored homozygous mutant (*Or*) [23] were finely powered in liquid nitrogen using a pestle and mortar, and suspended in 10 mL of extraction buffer (1% polyvinylpolypyrrolidone, 0.1 M sucrose, 0.1 M KCl, 0.1 M Tris (pH 7.5), 50 mM EDTA, 1 mM PMSF, and 10 mM DTT). The extraction was homogenized at 4°C for 30 min followed by centrifugation at 4000 rpm for 30 min. An equal volume of phenol was then added and the samples were rehomogenized and centrifuged at 10 000 × *g* for 30 min. The upper phenol phase was re-extracted twice with the above extraction buffer as reported previously [25]. Proteins were precipitated from the final phenol phase with five volumes of acetone at −20°C overnight and washed with cold methanol and acetone. The resulting protein pellets were dried and resuspended in 0.4 mL of pre-IEF solution (7 M urea/2 M thiourea/4% CHAPS). The protein concentration was determined by the Bradford assay using BSA as the standard [26].

2.3 2-DE and image analysis

The soluble crude cell extracts were run in the first dimension using a Multiphor II system (GE Healthcare, Piscataway, NJ, USA). IPG strips (24 cm) with nonlinear pH 3.0–10.0 gradients from GE Healthcare were used for the first dimension separation. Crude extract protein (150 μg) was mixed with IPG strip rehydration buffer (7 M urea, 2 M thiourea, 4% CHAPS, 10 mM DTT, and 2% IPG strip buffer). The solution (450 μL) was added to each lane of a rehydration tray and the strips were allowed to rehydrate overnight at room temperature. The IEF was carried out in a gradient mode increasing the voltage from 300 to 3500 V over 1.5 h followed by holding the voltage at 3500 V for 4.5 h. The temperature was kept at 20°C. After completion of the IEF run, the proteins were reduced and alkylated [27]. The strips were then transferred and apposed to 12% SDS-polyacrylamide vertical gels that were cast in-house at 24 × 21 cm using a DALT Six gel casting apparatus. The vertical gels were run using a DALT Six gel running system (GE Healthcare) at 20 mA for 30 min followed by 50 mA for 5–6 h until the bromophenol blue (BPB) front marker reached the bottom of the gel. The protein spots were visualized by colloidal Coomassie blue (CCB) staining using a NOVEX CCB staining kit (Invitrogen, Carlsbad, CA, USA) according to manufacturer protocols. The CCB-stained gels were scanned using a Typhoon 9400 laser scanner (GE Healthcare) and raw image files were analyzed by Progenesis software version 2005 (Nonlinear Dynamics, Newcastle-upon-Tyne, UK) [28, 29]. The average normalized volumes for each spot (% of total spot volume) from each sample group containing triplicate gel images were compared, and the spots with at least a two-fold differential expression between WT and *Or* samples were subjected to subsequent protein identification analysis. All the samples were analyzed in triplicate.

For the 1-D SDS-PAGE analysis of standard proteins, a mixture of 2 pmol BSA and 1 pmol lactoferrin were separated on a precast NOVEX 12% Tris/glycine mini-gel (Invitrogen) followed by CCB staining. For direct solution digestion of standard proteins, 100 μg of BSA and lactoferrin were individually digested with trypsin after reduction and alkylation [27].

2.4 In-gel digestion of excised gel spots

Following visualization of the gels, 46 differently expressed gel spots with variable staining intensity at different M_r and p*I* ranges across the gels were chosen and excised manually using an OneTouch spot picker (1.5 mm id). An additional spot (WT1274, shown in Fig. 1) was also excised for use in a spiking experiment (see below). The 1-D gel bands were excised using a scalpel and mixed prior to in-gel digestion. The subsequent in-gel digestion and tryptic peptide extraction were performed following a protocol as described by Shevchenko *et al.* [30] with slight modification. The gel pieces were washed and destained, seriatim with 50 μL of water, 50 μL of 50% ACN/50% 50 mM ammonium bicarbonate (pH 7.8), and 50 μL of 100% ACN. Once the samples were dried down completely, 0.2 μg of trypsin in 20 μL of 50 mM ammonium bicarbonate (pH 7.8)/10% ACN was added to each tube. The samples were left on ice for 15 min and incubated overnight at 37°C.

The supernatant was recovered, and the remaining peptides were then sequentially extracted from the gel in a series of solutions. The first was 50 μL of 50% ACN with 2% formic acid (FA) and the second was 50 μL of 90% ACN with 0.5% FA. For each extraction the samples were sonicated for 10 min before the supernatants were removed. All the supernatants were combined and dried in a SpeedVac (Thermo Savant, Holbrook, NY, USA).

2.5 MS analysis

Each sample was reconstituted in 3 μL of 50% ACN with 0.1% TFA prior to MS analysis and 1 μL was spotted on a MALDI target plate. Before the sample could dry, 0.5 μL of saturated matrix (10 mg/mL of CHCA in 50% ACN with 0.1% TFA and 1 mM ammonium phosphate) was spotted on top of each sample and allowed to dry completely. The samples were then subjected to MALDI MS/MS analysis using a 4700 Proteomics Analyzer equipped with TOF-TOF ion optics (Applied Biosystem) with 4700 Explorer version 2.0. The instrument was operated in 1 kV reflector positive ion mode and calibrated with a calibration kit (Applied Biosystems) containing a mixture of six standard peptides as a default calibration for spectra acquisition. The laser power was set to 4600 for MS and 5200 for MS/MS with CID off.

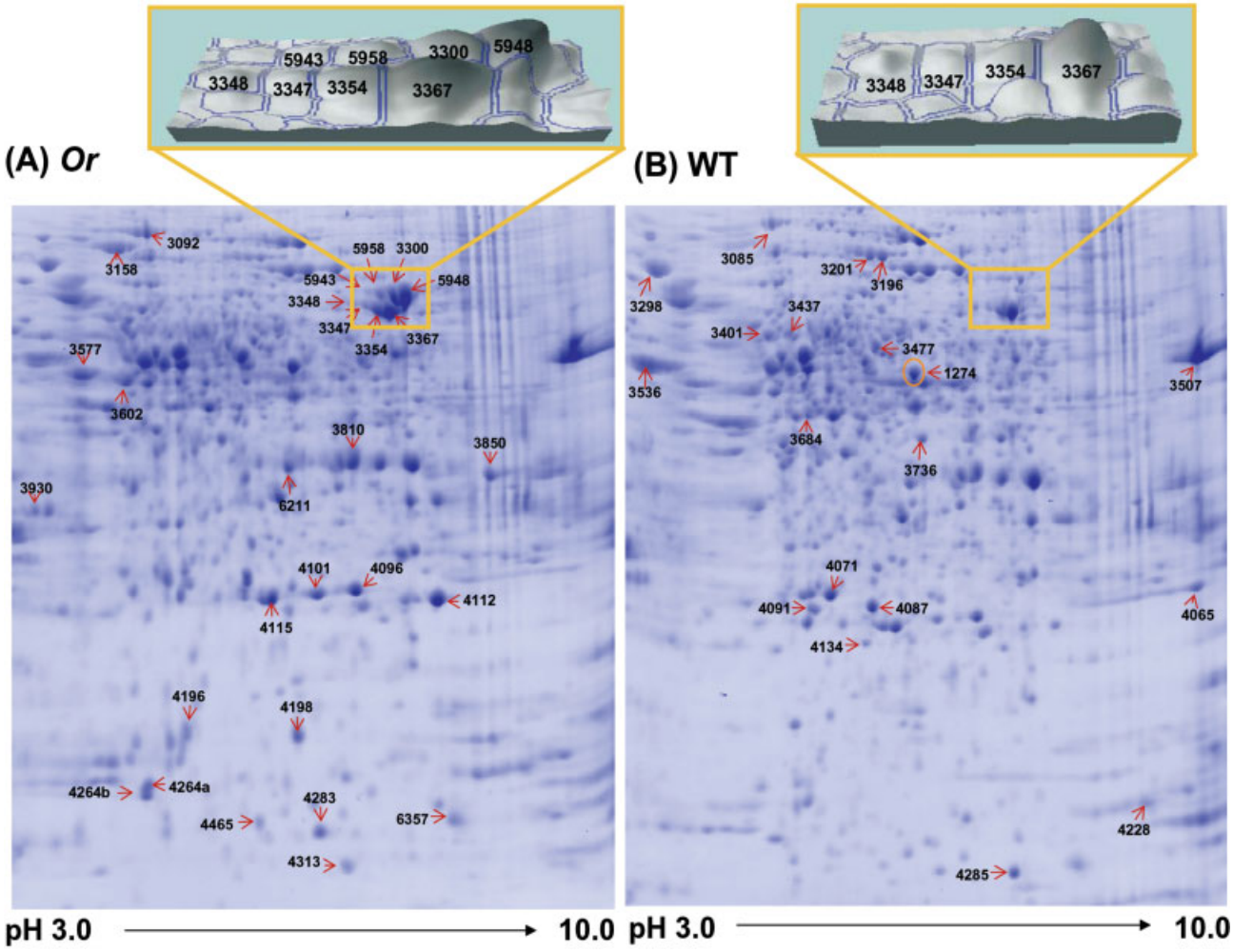


Figure 1. 2-DE gel images of 150 μg of protein extracts from edible curd tissue of (A) the cauliflower mutant (*Or*) and (B) WT cauliflower. The gels were CCB-stained and the spots with more than two-fold intensity difference selected for identification are noted by arrows and numbered. The circled spot WT1274 without differential intensity change was used for spiking experiment. An expanded view of eight gel spots in one area is highlighted with 3-D images analyzed by Progenesis software.

MS spectra were acquired across the mass range of 800–4000 Da with a minimum S/N filter at 25 for precursor ion selection. MS/MS spectra were acquired for the ten most abundant precursor ions with a total accumulation of 2000 laser shots.

The remainder of each 2-D spot sample was dried and reconstituted in 10 μL of 2% ACN with 0.1% FA for LC-ESI-MS/MS analysis. For analysis of direct solution digests, a mixture of 19 fmol BSA digest and 57 fmol lectoferrin digest (a ratio of 1:3) was used in quadruplicate. The extracted sample from the 1-D gel bands was reconstituted in 30 μL of 2% ACN with 0.1% FA and analyzed in triplicate. For the spiking experiment, the peptide extract from WT1274 was reconstituted in 12 μL of 2% ACN with 0.1% FA and split into four fractions. To each fraction, a mixture of the BSA/lactoferrin digests was spiked, at ratios of 5:1, 1:1, 1:2, and 1:5, respectively, prior to LC-MS/MS analysis.

The CapLC was carried out with an LC Packings Ultimate integrated capillary HPLC system equipped with a Switchos valve switching unit (Dionex, Sunnyvale, CA). The gel-extracted peptides (6.4 μL) were injected using a Famous autosampler (Dionex) onto a C18 PepMap trap column (5 μm, 300 μm × 5 mm, Dionex) for on-line desalting, and then separated on a PepMap C-18 RP capillary column (3 μm, 300 μm id × 150 mm, Dionex). A 30 min gradient of 5–45% ACN with 0.1% formic acid at 4 μL/min was used to elute the peptides. The CapLC was connected in-line to a hybrid triple quadrupole linear IT mass spectrometer, 4000 Q Trap (Applied Biosystems) equipped with a Turbo V source.

The data acquisition on the 4000 Q Trap was performed using Analyst 1.4 software (Applied Biosystems) in the positive ion mode for information-dependent acquisition (IDA) analysis. A 5 kV spray voltage was used for all the experiments. Nitrogen was used as both the curtain gas (value of

10 psi) and collision gas (set to high), and the heated interface was on. The declustering potential was set at 40 V to minimize in-source fragmentation. The ion source nebulizer gas was set to 25 psi. In an IDA analysis, after each survey scan from m/z 400 to 1600, an enhanced resolution scan was performed followed by MS/MS of the three highest intensity ions with multiple charge states. Rolling collision energy based on different charge states and m/z values was used to obtain optimal MS/MS spectra.

2.6 Data analysis and interpretation

The combined MS and MS/MS data from the MALDI-TOF/TOF and the LC-ESI-based IDA analysis were submitted to a MASCOT 1.9 for a database search against the NCBI nonredundant green plant database and mammals database (April, 2005). The search parameters allowed for one miscleavage, and variable modifications of methionine oxidation and cysteine carboxyamidomethylation with a mass tolerance of 75 ppm for MALDI, peptide tolerance = 2 Da, and MS/MS tolerance = 0.8 Da for LC-ESI. All the matches above a 95% confidence interval (CI) and only significant scores for the peptides defined by a MASCOT probability analysis (www.matrixscience.com/help/scoring_help.html#PBM) greater than "identity" were considered confidently-hit peptides and used for protein identifications.

The abundance of each identified protein by CapLC-MS/MS was estimated by determining the protein abundance index (PAI) [31] and the emPAI [22]. PAI is defined as the number of detected peptides divided by the number of observable peptides *per* protein normalized by the theoretical number of peptides expected *via in silico* digestion [31]. The emPAI is an exponential form of PAI minus 1 defined as $10^{\mathrm{PAI}}-1$ and the corresponding protein content in mole percent is calculated as mol-% = $\mathrm{emPAI}/\Sigma(\mathrm{emPAI}) \times 100$ [22]. The deviation factor was determined as described previously [22]. The BioAnalyst software (Applied Biosystem) was used to generate lists of *in silico* digested peptides for all the identified proteins.

3 Results and discussion

In comparative proteomics studies for complex samples, a quantitative 2-D gel approach typically requires the comparative analysis of numerous sets of gels in order to reveal different protein expression patterns across multiple experiments. This quantitative approach is based on the assumptions that the protein spots can be well resolved and that the changes in the normalized spot image intensity can be correlated with the expression level of particular proteins. However, as reported previously [16] and described here, multiple proteins are commonly identified from apparently single, well-resolved gel spots. Lim *et al.* [16] reported that up to 60% of 2-D gel spots analyzed by LC-MS/MS contained multiple proteins *per* spot. Even with a relatively narrower pH range of 4–7 for IEF, more than 20% of the analyzed spots have been found to contain multiple proteins [32]. Obviously, it is important to know how this observation affects the reliability of quantitative 2-D gel analysis. To better address and evaluate this issue, we purposely chose to use a relatively low resolution, broad IEF range (3–10) for our first dimension separation of a complex cauliflower protein extract to increase the frequency of comigrating proteins.

3.1 2-D gel and image analysis

The 2-D gel images of the extracted *Or* and WT cauliflowers are shown in Fig. 1. In general, the spots were well resolved, although some horizontal and vertical streaking can be observed at the basic end. This is a commonly observed phenomenon and is most often attributed to protein precipitation at basic pH, or the presence of a variety of PTMs. More than 1800 distinct protein spots were visualized with CCB stain which is consistent with our previous finding of >2700 spots when DIGE labeling was used (data not shown). In the subsequent image analysis using Progenesis Discovery software, one of the WT triplicate image gels was automatically selected as a reference gel. The triplicate gels for both WT and *Or* samples were averaged for image quantitative data analysis. The dynamic range of normalized intensity volumes cover from 0.011 to 1.181%, consistent with the expected two orders of magnitude detection range of CCB. Through this process 46 gel spots were identified as spots of interest because their expression level differed by more than a factor of 2 between WT and mutant. These were selected for further analysis and manually excised from the gel. They included 28 spots from the *Or* gel in which the protein expression was found to be higher than those of the WT (Fig. 1A) and 18 spots from WT for proteins that appeared to be expressed higher than those (Fig. 1B) in the mutant. All of the excised spots were digested and subjected to MS analyses as described above. The dynamic range covered by the normalized volumes for these 46 spots was 25-fold (0.045% for spot OR5943 to 1.087% for spot WT3507).

3.2 Protein identification by MS analysis

The 46 spots of interest from both *Or* and WT gels were analyzed by MALDI-TOF/TOF and CapLC-MS/MS analysis followed by database searching for protein identifications. Out of 46 spots, a confident identification of at least one protein was obtained for 43 using MALDI-TOF/TOF analysis and of these, 8 were found to contain two proteins. This produced a total of 51 proteins, including 38 unique proteins identified by MALDI-TOF/TOF analysis. Using CapLC-MS/MS, 44 spots were unambiguously identified, yielding a total of 108 proteins with 88 being unique. Interestingly, only one unique protein (gi2129934) was exclusively identified by MALDI-TOF/TOF analysis whereas a total of 50 unique proteins were identified by CapLC-MS/MS. A total of 38 unique proteins were identified by both approaches. Thus, a total of

89 proteins were identified from the 44 gel spots. Of the three spots that could not be identified by MALDI, only one (spot OR4313) was identified by subsequent analysis using CapLC-MS/MS. However, the basis of this identification was a single peptide hit (AQGDADSGVDR) from a 9.1 kDa protein (gi7630009, see Table 1). It should be noted that this small protein contains only two theoretically detectable tryptic peptides and each has a predicted grand average of hydropathicity (GRAVY) of −1.18 and −0.615, respectively, suggesting that both are very hydrophilic. The hydrophilic nature of the two peptides containing many residues with a negative mean desorption index [33] could possibly explain their absence in the MALDI-TOF/TOF analysis. Two spots (OR4196 and OR4198) could not be identified by either approach even though their intensities and normalized volumes were not significantly different from many of the identified proteins. There is no obvious explanation for this and since no meaningful MS spectra were found by manual

Table 1. Comparison of protein identifications for selected spots by MALDI-TOF/TOF and CapLC-MS/MS

Spot number	Accession number	Protein ID by LC-ESI-MS/MS	Protein ID by MALDI-TOF/TOF	M_r (kDa)	p*I*	MASCOT score	Pep-hit number	% Seq cov	emPAI	% Molar
OR3092	gi|1711296	Myrosinase binding protein		104.3	5.39	1500	25	37	3.11	100
			Myrosinase-binding protein	104.3	5.39	505	32	52		
OR3810	gi|21143	Unnamed protein product		37.0	7.70	601	11	38	6.85	64.1
			Unnamed protein product	37.0	7.70	438.9	15	58		
	gi|21592946	Fructose bisphos aldolase-like protein		38.5	6.05	555	9	29	3.83	35.9
			Fructose bisphosphate aldolase-like protein	38.5	6.05	316	14	35		
OR3158	gi|27372775	Lipoxygenase 2		101.4	5.27	887	16	20	2.26	92.7
			Lipoxygenase 2	101.4	5.27	458	15	19		
	gi|20259295	Unknown protein		67.5	5.30	138	3	5	0.18	7.3
OR4096	gi|3201613	Glutathione *S*-transferase		24.1	7.03	240	5	37	3.22	80.5
			Glutathione *S*-transferase	24.1	7.03	364	8	46		
	gi|19919694	Small Ran-related GTP-binding protein		25.1	6.66	149	6	34	0.585	14.6
	gi|600387	Proteosome subunit		25.3	7.82	63	1	5	0.194	4.9
WT3437	gi|99800	Chaperonin 62.5 K β-chain		62.4	6.56	1345	23	47	5.31	63.0
			Chaperonin 62.5 K β-chain	62.4	6.56	403	16	43		
	gi|58743500	Putative 2-isopropylmalate synthase		67.6	6.02	603	11	22	1.01	12.0
	gi|23297595	Putative phosphoglycerate dehydrogenase		63.3	6.16	497	10	15	2.01	23.8
	gi|9755778	Phosphoglucomutase-like protein		62.0	8.41	110	2	3	0.105	1.2
OR6211	gi|21593602	Putative malate dehydrogenase		35.5	6.11	631	14	43	12.11	87.2
			Putative malate dehydrogenase	35.5	6.11	423	11	55		
	gi|21592946	Fructose bisphosphate aldolase-like protein		38.5	6.05	271	5	17	0.833	6.0
	gi|21143	Unnamed protein product		36.9	7.70	260	5	20	0.585	4.2
	gi|6681337	Putative isocitrate dehydrogenase (NAD+)		40.6	6.71	76	1	4	0.145	1
	gi|23198416	Putative RNA-binding protein		41.3	6.08	57	1	3	0.212	1.5
OR4313	gi|7630009	40S ribosomal S21 homolog		9.1	7.79	66	1	13	2.16	100
			No hit							

inspection of the raw data files, it is unlikely that this was a consequence of the currently incomplete protein database for cauliflower. The list of protein identification results for 44 gel spots is shown in Supporting Information Table 1.

The combination of high mass accuracy, PMF data and MS/MS fragmentation offered by MALDI-TOF/TOF analysis makes it an ideal tool for rapid, sensitive and direct gel-based protein identifications. The MS/MS feature incorporated into this instrument uses fragmentation patterns for specific amino acid sequences to identify proteins and enhances the identification confidence over that which can be obtained from PMF data alone. Each of the identifications of the 51 proteins identified by MALDI-TOF/TOF was supported by at least one high-confidence MS/MS ion score. It was also found that five proteins were identified based on less than five tryptic peptides, but all were supported by at least 1 high confidence ion score.

Comparison of the protein identifications from the same gel spots by LC-MS/MS and MALDI approaches highlights a considerable difference in their capabilities as shown in Fig. 2. As many as six proteins were identified from a single spot by LC-MS/MS (Fig. 2A) whereas a maximum of two proteins were identified from a single spot using MALDI analysis (Fig. 2B). Furthermore, using the LC-MS/MS approach, 33 out of 44 spots (75%) were found to contain multiple proteins. Using MALDI analysis, only eight out of 43 spots (19%) were found to contain two proteins. This dramatic difference clearly demonstrates the advantages of coupling an LC separation to MS analysis, since it improves the dynamic range over that which can be obtained by direct MS analysis of 2-D gel-derived spots. The enhanced sensitivity of the LC-based IDA method is attributed to reduced ion suppression coupled with the enrichment of the separated peptides and online data-dependent acquisition analysis. Despite the many high-performance features incorporated into the MALDI-TOF/TOF instrument, the complexity of the typical "gel spot sample" is such that a direct analysis is frequently compromised. The use of an LC-MALDI-MS/MS analysis can be expected to improve this. A direct comparison of protein identifications by both MS methods for several selected spots containing two – five proteins is shown in Table 1.

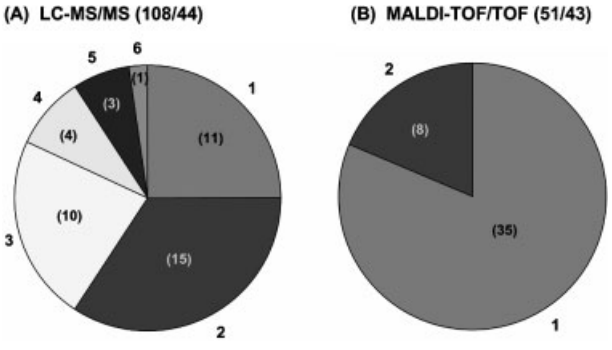


Figure 2. Comparison of the number of proteins identified from single spots by LC-MS/MS (A) and MALDI-TOF/TOF (B) approaches. Numbers in parentheses on inside of each segment are the number of spots found with the corresponding number of protein identifications *per* spot.

A further comparison of the protein identifications obtained by MALDI and LC-MS/MS of identical spots revealed that nearly all the proteins identified by MALDI were also identified by CapLC-MS/MS. Furthermore it was found that the top-ranked hit (highest MASCOT scores among the identified proteins in the spot) found in the MALDI analysis almost always matched the top-ranked hit found by LC-MS/MS. Table 2 shows that of the 43 spots identified by MALDI, 37 (86%) have the protein IDs that match the top-ranked hit of LC-MS/MS. For the remaining six spots (14%), the protein IDs obtained with MALDI matched the second-ranked hit obtained with LC-MS/MS. This result also provides evidence that direct MALDI-TOF/TOF analysis is comparable to LC-MS/MS identification of 2-D gel proteins, but lacks the capability to detect the less abundant proteins due to limited dynamic range. A similar conclusion can also be obtained through a comparison of the sequence coverage for the proteins identified by both MS approaches (data not shown).

3.3 Effects of multiple-protein identifications on 2-D gel image analysis

In order to accurately account for the presence of multiple proteins in a single gel spot and to evaluate the effect on the protein expression ratios observed in quantitative 2-D gel image analysis, it is necessary to find a way to establish the

Table 2. Comparison of the numbers of spots for protein IDs by MALDI to the hit-rank and abundance-rank determined by LC-MS/MS

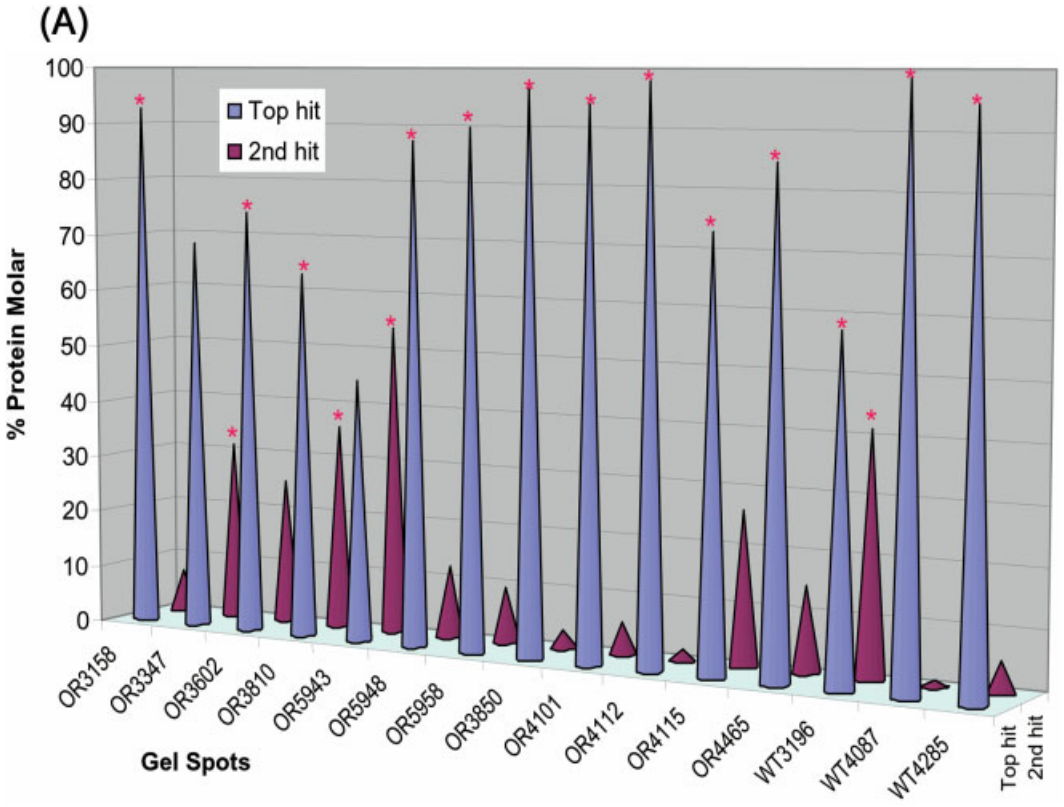
Items	Number of gel spots by MALDI-TOF/TOF analysis
Spots identified	43
Spots matched to top-hit of LC-MS/MS protein IDs	37 (86%)
Spots matched to second-hit of LC-MS/MS protein IDs	6 (14%)
Spots matched to the most abundant protein of LC-MS/MS[a)]	40 (93%)
Spots matched to the second-abundant protein of LC-MS/MS	2 (5%)
Spots matched to the third abundant protein of LC-MS/MS	1 (2%)

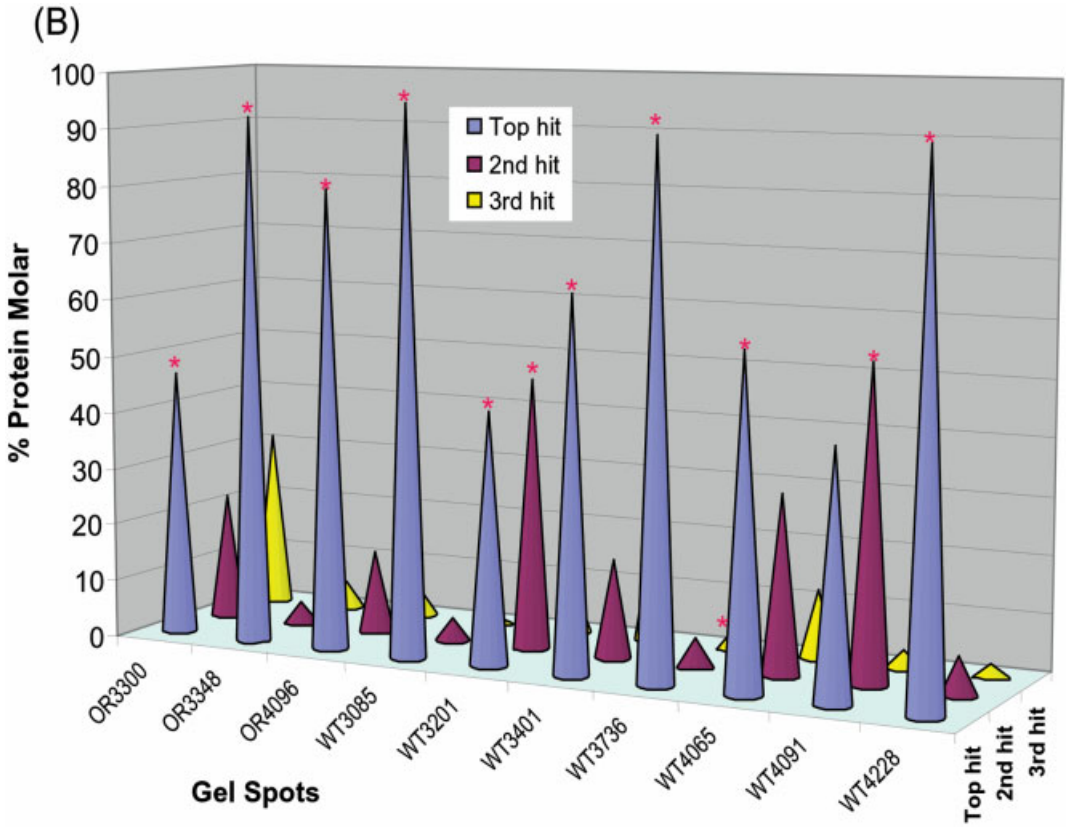
a) The abundance of each protein in multiprotein spots and its mole fraction percentage were determined based on emPAI as described in Section 3.

abundance of each of protein component of the gel spot and distribute the gel spot volume to each protein constituent appropriately. Based on many experimental observations it is known that several parameters such as hit rank (determined from probability scores), the number of detected peptides *per* protein (sequence coverage) and detected ion counts for the targeted peptides (signal intensity) can be correlated with protein abundance, although there is no compelling theoretical underpinning for the correlation. By normalizing these empirical parameters, Ishihama *et al*. have developed the emPAI that can be used to estimate the protein content in mole percent for complex protein samples [22, 31]. This concept was used in this experiment to determine the consequences of multiple-protein components within a single spot.

To validate the reliability of the emPAI approach used for quantitative estimation of protein contents, mixtures of two standard proteins (BSA and bovine lactoferrin) of known molar ratios were used in direct solution digests, in digests of 1-D gel bands followed by extraction and by spiking the standard mixture into a 2-D gel spot of cauliflower sample. The tryptic digests of BSA and lactoferrin at the ratio of 1:3 were initially analyzed by LC-MS/MS in quadruplicate. The experimental mole percentage obtained based on the detection of the peptides from BSA and lactoferrin are 23 and 77% with SD of 3.1%, which is in good agreement with the expected mole percent of 25 and 75. The digests of 1-D gel bands at the ratio of 2:1 for BSA to lactoferrin was analyzed in triplicate. The results showed that the mole percentage of BSA and lactoferrin were 68 and 32 with SD of 4.1%, demonstrating that the emPAI approach for estimation of protein abundance works equally well for the gel-based protein mixtures compared to the direct solution digest samples. To further evaluate the feasibility of using the emPAI approach for 2-D gel spots of the cauliflower samples used in this study, a control spot (WT1274, as shown in Fig. 1) with relatively high stain intensity was selected for emPAI analysis spiked with the BSA and lactoferrin digests. The aliquots of extracted tryptic peptides from the 2-D gel spot were spiked with the mixture of BSA and lactoferrin digests at a ratio of 5:1, 1:1, 1:2, and 1:5, respectively. The results of these experiments yielded two protein identifications consistent with the plant-based nature of the sample (enolase and adenosylhomocysteinase) at mole percent of 85 and 15 (calculated exclusive of the spiked proteins) with an SD of 1.7% regardless of the amount of spiked digests added. The mole percent of the spiked proteins consistently showed good agreement with the expected ratios of BSA to lactoferrin (77–23, 48–52, 39–61, and 10–90, observed *versus* 83–17, 50–50, 33–67, and 17–83, expected). The details of detected peptides and emPAI analysis for the solution, gel-based and spiked 2-D gel samples are shown in Supporting Information Table 2. This preliminary data demonstrate that the emPAI approach for estimation of protein abundance for 2-D gel spots is feasible, reliable, and relatively accurate.

The protein abundance in all 44 spots identified by CapLC-MS/MS was then determined using the emPAI as described above. All calculated PAI, emPAI values, and their mole fraction percentages for the 108 identified proteins in the 44 spots investigated are listed in Supporting Information Table 1. Figure 3 shows the comparison of protein content distribution for the 33 spots containing multiple proteins calculated based on the emPAI. In Fig. 3A, 15 distinct spots containing two proteins are shown and the relative abundance of the two proteins is compared based on individual emPAI values. The results show that the top-hit proteins are the most abundant contributing on average 81.5 $\pm$ 16.4% of the spot volume. Spot OR3810 and WT3196 contain relatively high abundance for second ranked proteins with 36 and 42%, respectively (Fig. 3A). MALDI-TOF/TOF analyses for both of these spots were found to contain the same two proteins identified by LC-MS/MS which is indicated by the red star above the cones. After adjustment by the emPAI, spot OR5943 shows slightly higher abundance for second ranked protein (54%) than for the top-ranked protein (46%). Interestingly, this second ranked protein hit was that identified by the MALDI analysis as a single protein hit for this spot. Figures 3B and C show the ten spots having three protein IDs and eight spots with four–six protein IDs, respectively. As was observed above, the top-hit proteins were found to be the most abundant proteins with very few exceptions. The identification determined by CapLC-MS/MS for the five spots (WT3201, WT3736, WT3437, WT3477, and WT3684) containing two proteins that were consistently matched to the identifications determined by MALDI. Table 2 gives the detailed information concerning the number of MALDI IDs that match the top-ranked hits (determined by LC-MS/MS) and the most abundant proteins (determined by the emPAI on LC-MS/MS data). The results show that Protein IDs from 37 out of 43 spots (86%) identified by MALDI match the top-ranked protein IDs from LC-MS/MS. The protein IDs from the remaining six spots (14%) match the number 2 ranked protein IDs as determined LC-MS/MS. Thus, it is particularly interesting to further investigate the relationship between the hit rank and the estimated relative abundance of the identified proteins for those six spots. Figure 4 shows the relative protein abundance of the identified proteins *per* spot for all six spots where protein IDs by MALDI match to the second hit protein IDs of the LC-MS/MS approach. It was found that three (OR5943, WT4091, and OR3354) of the six spots show slightly higher relative abundance for the second ranked hit protein (~50%) over the top-ranked hit protein (~40%). This observation further demonstrates bias in MALDI toward identification of the most abundant protein. It is not clear why in the remaining three spots (OR3347, OR3577, and OR3930) MALDI appears to identify the second or even the third most abundant protein (in one case). It is possible that the nature of these proteins is such that they are difficult to analyze by MALDI. However, upon investigation all three top-hit proteins by LC-MS/MS appear to possess the tryptic peptides containing





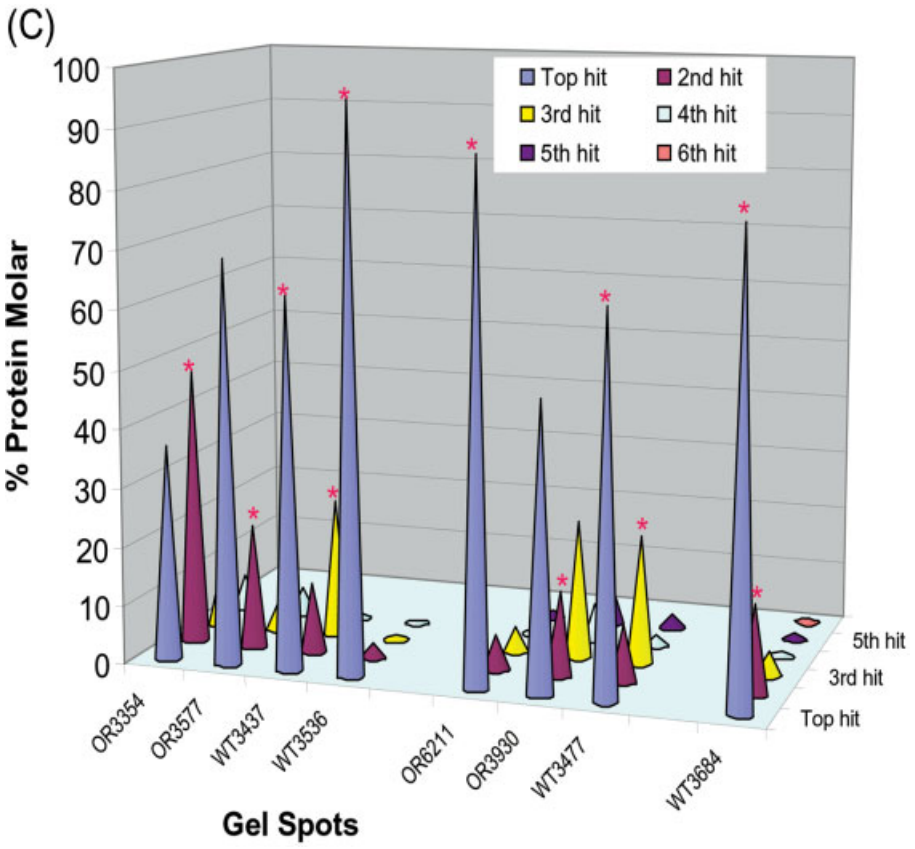


Figure 3. Comparison of percent protein content for each protein in spots containing multiple proteins based on the emPAI: (A) for the 15 individual spots containing two proteins; (B) for the ten individual spots containing three proteins, and (C) for eight individual spots having four–six proteins. The mole percentages for top-hit, second-hit, and third-hit proteins are indicated by blue, dark-red, and yellow cones, respectively. The red * label above the cones denotes the proteins were also successfully identified by MALDI-TOF/TOF.

"normal"amino acid composition with 6–7% arginine residue [34] and M_r. As shown in Table 2, after matching the numbers of spots identified by MALDI to the relative abundance rank by LC-MS/MS, the protein IDs from 93% (40/43) of the spots identified by MALDI indeed matched the protein with the highest abundance. This result indicates that the estimated protein abundance based on the emPAI calculation for multiple-protein identifications in single spots appears to be reliable and does reflect a reasonably quantitative relationship.

To estimate the contribution of the top-hit protein in the spots found to contain multiple proteins to the spot intensity determined by 2-D gel image analysis, two important factors were integrated: the average percent abundance of the top-hit proteins and the size (weight) of fractions for the number of identified proteins in a single spot. Table 3 shows the results of weighted top-hit protein contributions to the corresponding spot image intensity for the 44 top-hit proteins by LC-MS/MS analysis. It appears that the weighted fractions of multiple-protein IDs together with the average percent protein abundance can be used to estimate the contribution from the top-hit protein to the observed change in spot volume (determined by 2-D gel image analysis) for any gel spot. The result indicates that the weighted average of the top-hit proteins contributes over 81% of the gel spot volume for the 44 gel spots. In other words, the average contribution to the gel spot volume from all the other protein components is nearly 20%. Whereas the average of molar percentage for the 43 top-hit proteins identified by direct MALDI analysis is 79% (Supporting Information Table 1). This indicates that by using the usual direct analysis workflow, 20% of the staining intensity (on average) is incorrectly assigned to the most abundant protein in the spot. Based on these observations, we believe that assigning the change in spot volume determined by 2-D image analysis to the top-hit (most abundant) protein identified by MALDI is still reasonable, provided the fold change criteria for determining significant changes is large (*e.g.*, >two-fold) and the mole fraction of the most abundant protein is much greater than that of any of the minor components. As we have limited spot numbers (44) studied in this work, it is difficult to suggest a general fold change criterion for use in other studies. However, we did use a low-resolving first-dimensional IPG strip (three–ten) which represents something of a "worst case" scenario. Even under these conditions with 75% of the analyzed spots having multiple proteins identified, the impact of minor component comigrating proteins is, on average, still limited. Clearly, the impact of comigration of proteins can be reduced further by increasing the potential resolving power of the gel, such as by using narrow pH range strips for IEF. Table 3 also

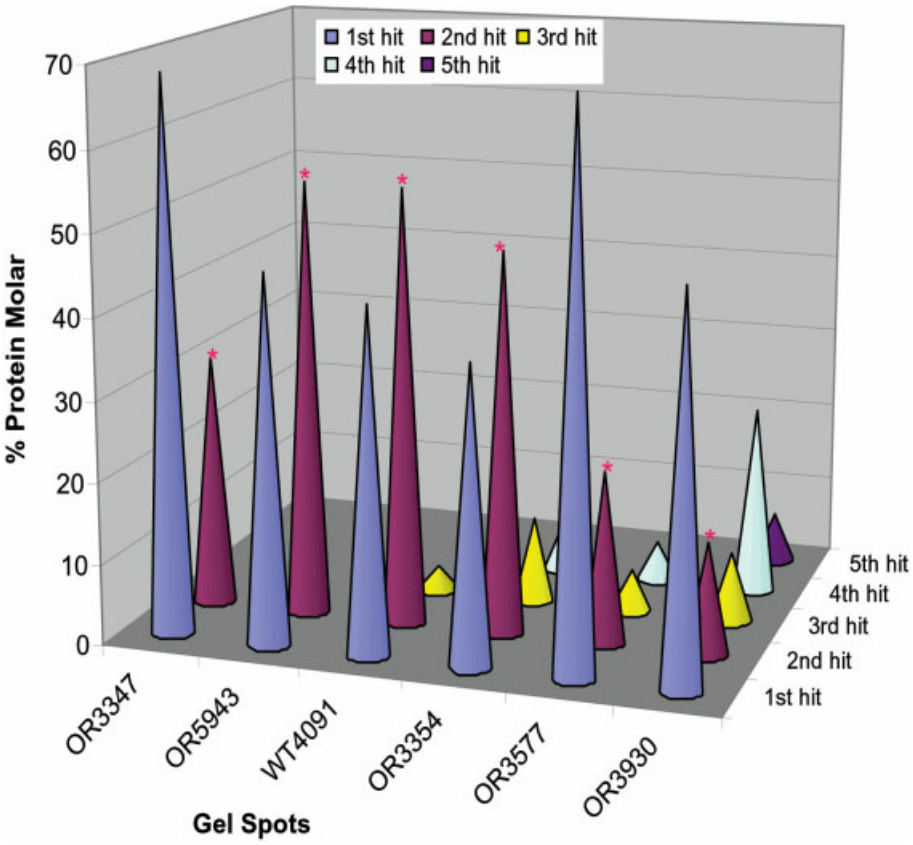


Figure 4. Comparison of protein mole percentages for six single spots where the MALDI-TOF/TOF-identified proteins for each spot matched the second-ranked identification by LC-MS/MS. The mole fraction percentages for top-hit and second-hit proteins are indicated by blue and dark-red cones, respectively. The red * label above the cones denotes the protein identified by MALDI in the corresponding spots.

shows the results of weighted contributions to the gel spot staining intensities for the 33 spots containing multiple proteins. These data can be used to estimate the overall top-hit proteins for multiprotein spots, excluding single protein IDs for each spot. The result indicates that the top-hit protein in spots containing multiple proteins contributes about 75% of the gel spot staining intensities, which is still acceptable providing a large fold change criteria for determining significant differences is used.

To see the correlation between the normalized volume (staining intensities) of 44 gel spots from Progenesis image analysis and their emPAI values for the 108 identified proteins, we plotted the normalized volumes (representing the total protein concentration in the spots) against the emPAI values (representing the abundance of specific proteins) calculated from the identified peptides for both the top-hit proteins and the combination of all the proteins in the spots. This can provide additional validation of the accuracy of protein abundance calculations. Figure 5 shows the plot of normalized spot intensities from 44 spots in this work against the individual emPAI values of 44 top-hit proteins (indicated by diamonds) and the combined emPAI values for all the identified proteins in each of the spots (indicated by squares). The results show a linear relationship for the top-hit proteins with correlation coefficient $r = 0.66$, deviation factor (average $\pm$ SD) = 2.77 $\pm$ 3.08, and for the combined proteins with $r = 0.69$, deviation factor (average $\pm$ SD) = 2.23 $\pm$ 2.45. This overall linearity demonstrates that the experimentally obtained protein abundance estimation is well correlated with the normalized spot intensities. The fact that the plot for the combined proteins yields only slightly better correlation coefficient and deviation factor than those of top-hit proteins, also provides additional evidence that the other comigrated proteins have a small impact on the interpretation of the image analysis data. It should be noted that there are six outlier spots (WT3507, OR5948, OR3577, OR3367, OR4115, and OR3300), whose top-hit protein and combined emPAIs were not included in Fig. 5. As shown in Fig. 1, all six spots show larger staining area and possess extremely high intensities with normalized volume from 0.5 to 1.1%. This study used manual spot picking with an OneTouch manual spot picker (1.5 mm id) and so most of the stained proteins in these spots were not excised. It is therefore not surprising that those spots were outliers with much lower emPAI values against the detected normalized intensity volumes directly imported from software analysis.

The PAI and emPAI tools were developed for their use in fully sequenced genomes and their suitability for studying proteins from a nonsequenced genome has not yet been confirmed. If a spot of interest contains a high proportion of a nonidentifiable protein (one for which either there is no entry in the database or being heavily modified by PTMs or

Table 3. Weighted top-hit protein contributions in total 44 gel spots to the corresponding gel spot intensities

Proteins/spot	Spot numbers	Average mol-% of top-hit protein	Weighted contributions (%) for total spots	Weighted contributions (%) for multiprotein spots
1	11	100	25.0 (11/44 × 100)	
2	15	81.5	27.8 (15/44 × 81.5)	37.0 (15/33 × 81.5)
3	10	71.0	16.1 (10/44 × 71.0)	21.5 (10/33 × 71.0)
4	4	66.0	6.0 (4/44 × 66.0)	8.0 (4/33 × 66.0)
5	3	66.3	4.5 (3/44 × 66.3)	6.0 (3/33 × 66.3)
6	1	78.1	1.8 (1/44 × 78.1)	2.4 (1/33 × 78.1)
Overall	44		81.2[a)]	74.9[b)]

a) Calculations of the weighted contributions to the gel spot staining intensities for the 44 top-hit proteins by LC-MS/MS indicates that, overall, the top-hit proteins in any spots contribute over 81% of the gel spot staining intensities. Therefore, the average contribution from the minor protein components of the spot is less than 20%.

b) Calculations of weighted contributions to the gel spot staining intensities for the identified 33 spots indicate that the overall top-hit protein in multiprotein spots contribute about 75% of the gel spot staining intensities.

containing many nonisobaric amino acid substitutions) the normalized spot volume contributed by this protein could be falsely assigned to the identifiable proteins found within the spot. Such a situation would be expected to generate significant numbers of interpretable mass spectra that yield no protein identification after the database search. If such a situation arises then the mole fraction information generated by the emPAI analysis should be viewed as suspect. We note that no such phenomenon in the results reported here. Although the genome of cauliflower has not been fully sequenced, there are significant amounts of *Brassica* sequence information. In addition, cauliflower and *Arabidopsis* typically share over 85% nucleotide sequence identity in coding regions [35–37]. Therefore, it is not surprising that using the green plant database (including *Arabidopsis*) as was done for this study, we did not find even small numbers of good mass spectra that failed to result in an identification.

3.4 Influence of the integrated approach on gel-based comparative proteomics

In an attempt to evaluate the impact of multiple proteins in single spots on comparative proteomics by conventional 2-D Gel/MALDI approach, we have developed an integrated approach that combines 2-DE, image analysis, LC-MS/MS analysis of 2-D gel spots and emPAI-based quantitative determination of identified proteins. This integrated approach is called "2-D GeLC-MS/MS" analysis reflecting the combination of 2-D gel separation and subsequent LC-MS/MS analysis, which distinguishes from the traditional 1-D gel-based GeLC-MS/MS approach [24]. In this 2-D GeLC-MS/MS workflow, complex protein extracts are first separated on a broad p*I* range high-performance 2-D gel followed by LC-MS/MS analysis of tryptic digests. The results obtained using the approach demonstrate that this is an effective strategy for the improved reliability of the current quantitative proteomics analysis. Changes in spot intensity (spot volume) are determined by staining with an appropriate protein stain followed by image analysis. The protein components in spots from the gels that are being compared can be determined through LC-MS/MS analysis and their mole fractions calculated through emPAI. A portion of the total spot volume (staining intensity) can then be assigned to each protein component on the basis of its mole fraction. By comparing these fractional spot volumes for each protein component between samples, the changes in protein abundance can be directly determined. Using this approach the total protein concentration represented by a spot is determined by its staining intensity and the relative concentration of the individual protein components of the spot are determined on the basis of their mole fraction.

The results generated from this novel approach in this work suggest that the average contribution of the top-hit protein identified by LC-MS/MS and MALDI to the gel spot staining intensity is approximately 80%. Thus, assigning the observed change in staining intensity to the most abundant proteins or top-hit proteins is generally reasonable and that the impact of the low abundance proteins appears limited. However, given the fact that for 7% of spots analyzed by MALDI the top-hit protein was not found to be the most abundant protein as determined by the emPIA and that only 19% of spots are identified by MALDI as multiple-protein spots *versus* 75% by LC-MS/MS, we believe that an alternative method incorporating an additional dimension of separation to determine the change in expression of the minor components is needed. As more and more of the highly abundant proteins are characterized in proteomics analysis, the need for targeting the changes of low abundant proteins will increase in near future. Therefore, the integrated 2-D GeLC-MS/MS approach presented here is extremely useful and important because it allows for the quantitative measurement of those relatively low abundant proteins which are often not detected by direct MALDI analysis or are identified as minor, low abundant components. The combined

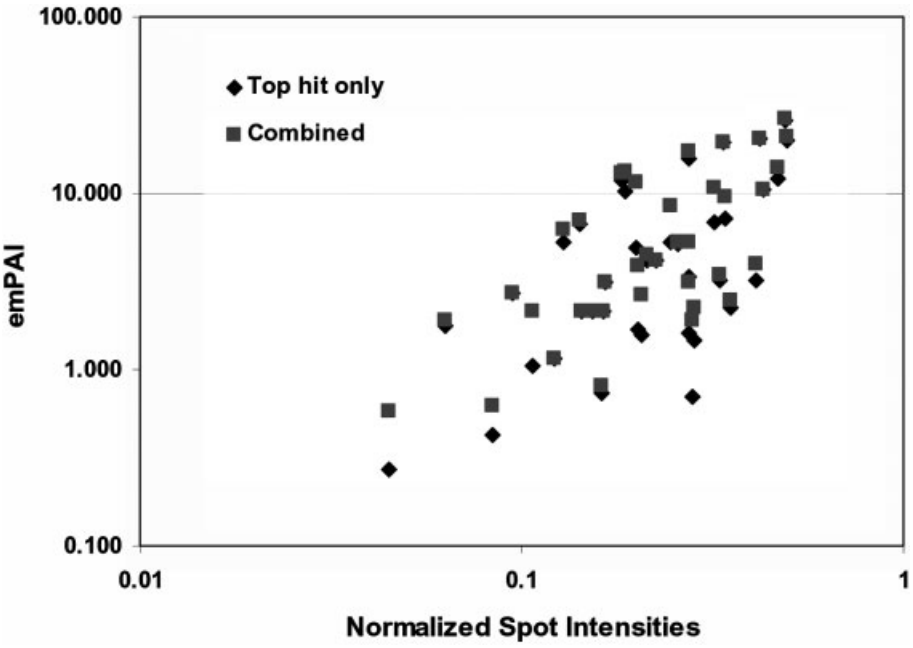


Figure 5. Relationship between normalized staining intensities of 44 gel spots and their emPAI values for identified proteins. Diamonds indicate the calculated emPAI values for top-hit protein only while squares denote the combined emPAI values for all the proteins identified in the single spots excluding six outlier spots (see Section 3).

approach developed in this work is unique in that it can be used to assign the change is expression for each of the protein components of a spot directly, even the low abundance ones. Thus, it represents a significant advancement for gel-based proteomics. It should be note that the limitations of this 2-D GeLC-MS/MS quantitative approach include the dependence on an accurate determination of the emPAI in which uninterpretable MS spectra may cause considerable error (as described above). In addition, the two equivalent spots from both the upregulated and downregulated gels have to be analyzed, which will double the number of samples to be analyzed, the time needed to complete the analysis, and the associated costs.

Recently, there have been several gel-free quantitative techniques developed that are currently being used successfully. These include ICAT [38], O^{18} labeling [39], stable isotope-labeled amino acid culture (SILAC) [40], isobaric tags for relative and absolute protein quantitation (iTRAQ) [41] and isotope-coded protein labeling (ICPL) [42]. All such methods are based on pre-/postisolation chemical labeling of proteins/tryptic peptides, in solution followed by traditional shotgun analysis using multidimensional LC-MS and LC-MS/MS. Therefore, these solution-based methods can overcome some of the limitations of gel-based analysis such as detecting very hydrophobic proteins, extremely basic or acidic proteins, extremely large or small proteins. Furthermore these methods can improve reproducibility (when compared to single stain gel analysis) and can provide higher throughput quantitative analysis. However, there are some limitations of these gel free methods. For example, the labeling chemistries are usually sensitive, condition-specific and subject to interference by sample impurities and matrix effects. In addition, each of these methods has its own set of weakness such as SILAC is not applicable to all proteomics experiments and ICAT only targets cysteine containing proteins. The higher resolution protein separation attainable by 2-DE allows for quantitative identification of relatively low abundance proteins and structural characterization of differentially processed or post-translationally modified forms of a protein. This is a key advantage of gel-based quantitative analysis. 2-D GeLC-MS/MS/emPAI provides researchers a methodology to deal with one of the key disadvantages of gel-based quantitative analysis, the comigration of proteins. We believe that by eliminating this disadvantage it will prove to be an alternative and complement to gel-free methods.

4 Concluding remarks

Many instances were found in which apparently well-resolved protein spots gave rise to multiple-protein identifications and 75% of the 2-D gel spots analyzed were found to contain multiple proteins in this study. Through comparison of 2-D gel-based protein identifications by both CapLC-MS/MS and MALDI-TOF/TOF, we have thoroughly evaluated the effect of multiple-protein identifications from a single spot on its gel image intensity analysis. In comparison of protein identifications of 2-D gel spots by direct MALDI-TOF/TOF and LC-based ESI-MS/MS, we have confirmed that the LC-MS/MS yields a significantly improved dynamic range, enabling the MS/MS detection of low abundance proteins present in a spot.

In comparing protein identifications by hit-rank from LC-MS/MS and MALDI, it was found that 86% (37 out of 43) of the spots identified by MALDI matched the top-hit protein identified by LC-MS/MS. Estimation of the relative

protein abundance of all 33 spots found to contain multiple proteins and the 11 single protein spots using emPAI indicates that 40 out of 43 spots identified by MALDI matched the most abundant protein IDs from LC/MS/MS. It was found that the weighted average of the 44 top-hit proteins contributes over 81% of the gel spot staining intensities for the 44 gel spots. Therefore, the contribution, on average, from the minor protein components of the spot is less than 20%. The normalized spot intensities correlate well with the experimentally obtained protein abundance determinations for both top-hit proteins and combined multiple proteins in the spots. The results presented in this study demonstrate that the gel-based image analysis for protein quantitation of top-hit proteins is generally reliable; nevertheless, great care must be taken when interpreting data particularly in the case that the observed fold change is low. It was demonstrated that direct MALDI analysis of 2-D gel spots has a bias for the detection of the most abundant proteins but the impact of minor component, comigrating proteins is slight, particularly when studying highly abundant proteins and when the fold change criteria used to determine significant differences is high.

The results presented here also suggest that by using direct MS analysis (by MALDI or any other ionization technique) of 2-D gel separated proteins, a great deal of the information is missed due to limited dynamic range of direct analysis. Adding an LC separation prior to MS/MS analysis greatly increases the number of gel spots that are found to contain multiple proteins, and it is possible to determine the mole fraction of all the proteins detected within the spot by the emPAI. These observations have lead us to develop an integrated 2-D GeLC-MS/MS approach which combines 2-D gel-based LC-MS/MS with emPAI analysis to correctly assign the change in a spots staining intensity to each of the component proteins found in a gel spot on the basis of its mole fraction. This integrated workflow has been extremely useful, allowing for quantitative measurement of the average rate of false assignment using direct MALDI analysis and quantitative measurement of low abundant proteins, which are often not detected by direct MALDI analysis. The 2-D GeLC-MS/MS/emPAI method is a complementary technology to the existing gel-free quantitative techniques for comparative proteomics study.

We would like to thank Professor Jocelyn Rose (Cornell University), Drs. Colleen van Pelt, Xiaolong Yang, and Kevin Howe for critical reviewing of this manuscripts and helpful discussions. We gratefully acknowledge the USDA-ARS for its generous support. This work was also supported in part by USDA-NRI grant 2001-03514.

5 References

[1] Aebersold, R., Mann, M., *Nature* 2003, *422*, 198–207.

[2] Aebersold, R., *J. Am. Soc. Mass Spectrom.* 2003, *14*, 685–695.

[3] Peng, J., Gygi, S. P., *J. Mass Spectrom.* 2001, *36*, 1083–1091.

[4] Gygi, S. P., Aebersold, R., *Curr. Opin. Chem. Biol.* 2000, *4*, 489–494.

[5] Celis, J. E., Gromov, P., *Curr. Opin. Biotechnol.* 1999, *10*, 16–21.

[6] Giavalisco, P., Nordhoff, E., Kreitler, T., Kloppel, K. D. *et al.*, *Proteomics* 2005, *5*, 1902–1913.

[7] Rabilloud, T., *Proteomics* 2002, *2*, 3–10.

[8] Pluskal, M. G., Bogdanova, A., Lopez, M., Gutierrez, S., Pitt, A. M., *Proteomics* 2002, *2*, 145–150.

[9] Lopez, M. F., *J. Chromatogr. B* 1999, *722*, 191–202.

[10] Washburn, M. P., Wolters, D., Yates III, J. R., *Nat. Biotechnol.* 2001, *19*, 242–247.

[11] Link, A. J., Eng, J., Schieltz, D. M., Carmack, E. *et al.*, *Nat. Biotechnol.* 1999, *17*, 676–682.

[12] Lin, D., Tabb, D. L., Yates, J. R., *Biochim. Biophys. Acta* 2003, *1646*, 1–10.

[13] Thiede, B., Hohenwarter, W., Krah, A., Mattow, J. *et al.*, *Methods* 2005, *35*, 237–247.

[14] Vanrobaeys, F., Van Coster, R., Dhondt, G., Devreese, B., Van Beeumen, J., *J. Proteome Res.* 2005, *4*, 2283–2293.

[15] Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S. *et al.*, *Nature* 1996, *379*, 466–469.

[16] Lim, H., Eng, J., Yates, J. R., III, Tollaksen, S. L. *et al.*, *J. Am. Soc. Mass Spectrom.* 2003, *14*, 957–970.

[17] Gorg, A., Weiss, W., Dunn, M. J., *Proteomics* 2004, *4*, 3665–3685.

[18] Gharbi, S., Gaffney, P., Yang, A., Zvelebil, M. J. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 91–98.

[19] Unlu, M., Morgan, M. E., Minden, J. S., *Electrophoresis* 1997, *18*, 2071–2077.

[20] Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., Aebersold, R., *Proc. Natl. Acad. Sci. USA* 2000, *97*, 9390–9395.

[21] Zhu, H., Bilgin, M., Snyder, M., *Annu. Rev. Biochem.* 2003, *72*, 783–812.

[22] Ishihama, Y., Oda, Y., Tabata, T., Sato, T. *et al.*, *Mol. Cell Proteomics* 2005, *4*, 1265–1272.

[23] Li, L., Paolillo, D. J., Parthasarathy, M. V., Dimuzio, E. M., Garvin, D. F., *Plant J.* 2001, *26*, 59–67.

[24] Schirle, M., Heurtier, M. A., Kuster, B., *Mol. Cell. Proteomics* 2003, *2*, 1297–1305.

[25] Saravanan, R. S., Rose, J. K., *Proteomics* 2004, *4*, 2522–2532.

[26] Bradford, M. M., *Anal. Biochem.* 1976, *72*, 248–254.

[27] Zhang, S., Van Pelt, C. K., Henion, J. D., *Electrophoresis* 2003, *24*, 3620–3632.

[28] Rosengren, A. T., Salmi, J. M., Aittokallio, T., Westerholm, J. *et al.*, *Proteomics* 2003, *3*, 1936–1946.

[29] Marengo, E., Robotti, E., Antonucci, F., Cecconi, D. *et al.*, *Proteomics* 2005, *5*, 654–666.

[30] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, *68*, 850–858.

[31] Rappsilber, J., Ryder, U., Lamond, A. I., Mann, M., *Genome Res.* 2002, *12*, 1231–1245.

[32] Link, A. J., Hays, L. G., Carmack, E. B., Yates III, J. R., *Electrophoresis* 1997, *18*, 1314–1334.

[33] Valero, M. L., Giralt, E., Andreu, D., *Lett. Pept. Sci.* 1999, *6*, 109–115.

[34] Krause, E., Wenschuh, H., Jungblut, P. R., *Anal. Chem.* 1999, *71*, 4160–4165.

[35] Babula, D., Kaczmarek, M., Barakat, A., Delseny, M. *et al.*, *Mol. Genet. Genomics* 2003, *268*, 656–665.

[36] Cavell, A. C., Lydiate, D. J., Parkin, I. A., Dean, C., Trick, M., *Genome* 1998, *41*, 62–69.

[37] Lan, T. H., DelMonte, T. A., Reischmann, K. P., Hyman, J. *et al.*, *Genome Res.* 2000, *10*, 776–788.

[38] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F. *et al.*, *Nat. Biotechnol.* 1999, *17*, 994–999.

[39] Yao, X., Freas, A., Ramirez, J., Demirev, P. A., Fenselau, C., *Anal. Chem.* 2001, *73*, 2836–2842.

[40] Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 376–386.

[41] Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 1154–1169.

[42] Schmidt, A., Kellermann, J., Lottspeich, F., *Proteomics* 2005, *5*, 4–15.

